# Chapter 4
# Linking Low Resolution Protein Folds to Biochemical Relevance: Accurate Protein Structure, Dynamics and Thermodynamics

As we move to an era of genetic information at the level of complete genomes, classifying the fold topology of each sequence in the genome is a vital first step toward understanding gene function. However, the ultimate limitation in fold recognition is that these algorithms only provide "low-resolution" structures. It is crucial to enhance and develop methods that permit a quantitative description of protein structure, dynamics and thermo-dynamics, in order relate specific sequence changes to structural changes, and structural changes to associated functional/phenotypic change.

For example, these more accurate approaches will greatly improve our ability to modify proteins for novel uses such as to change the catalytic specificity of enzymes and have them degrade harmful waste products. While the tertiary fold of proteins involved in disease presumably remain invariant under mutation, quantitative differences in structure between wild type and mutant can have important macroscopic effects on function that can be manifested as disease. More accurate screening for new drug targets that bind tightly to specific protein receptors for inhibition will require quantitative modeling of protein/drug interactions. Therefore the next step is the quantitative determination of protein structure starting from the fold prediction, and ultimately directly from sequence.

***Key milestones to be addressed during the Strategic Simulation Initiative:***

- ***Continued development and verification of protein and aqueous force fields***
- ***Development of robust optimization/search methods to construct biochemically relevant protein structures from the endpoint of fold recognition algorithms***
- ***Significant enhancement in the range of protein topologies and fold types that have characterized folding free energy landscapes.***
- ***Resources for long timescale dynamics and thermodynamic simulations of protein folding and function.***

Addressing these milestones will present challenges on multiple levels and necessitate integration across scientific disciplines from biology, physics and chemistry to mathematics and computer and informational sciences. The computational complexity presented by simulation problems posed for the determination of protein structure, exploration of folding landscapes, and protein function will outstrip current computational infrastructure and require “many teraflop” architectures, distributed data storage, archival and management systems and software developments in simulation methodologies as well as scheduling and distributed computing paradigms.

# A Global Optimization Strategy for Refining Protein Structure From Fold Assignment

Once a new sequence is correctly assigned and aligned to a target fold, a large number of constraints are imposed by the fold topology. These include α-helical and β-sheet structure that can be manifested as soft constraints to use within both a local optimization algorithm and as guidance within various global optimization frameworks.

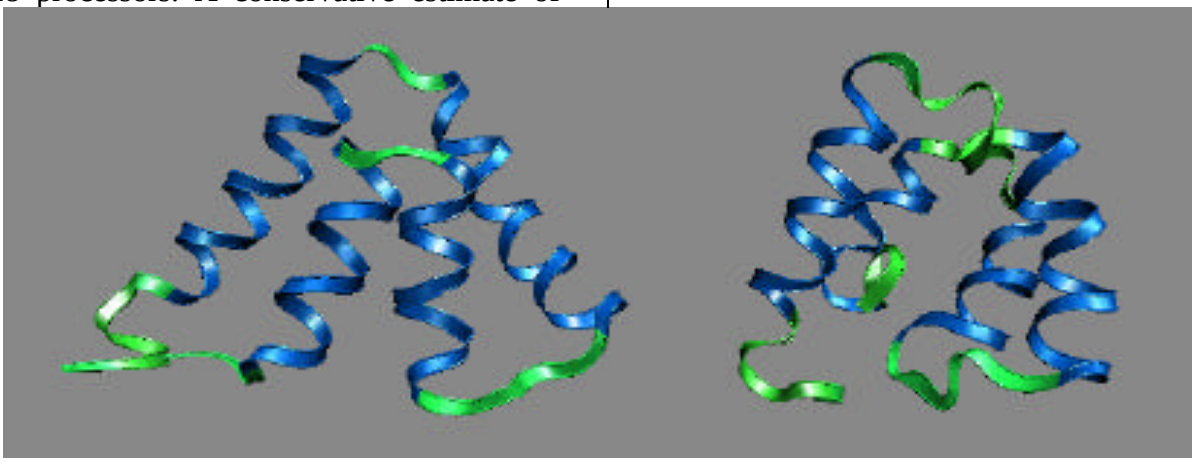
The use of soft constraints permit partial solution to the global optimization problem within a local optimization context. It also bridges the gap between primary and tertiary structure by greatly narrowing the search space, by focusing the global optimization work on sequence regions predicted to be coil or loop, regions where proteins exhibit the most variation in structure.

The figure below shows the prediction of a 70 amino acid segment of a helical protein, uteroglobin, using such a global optimization strategy. Soft-constraints were defined from neural network calculations, and helical regions were formed with locally constrained minimizations. Regions of the sequence predicted to be coil were subject to global optimization based on sampling and perturbation. The developed strategy was run on the T3E at NERSC using between 16 and 128 processors. A conservative estimate of the number of FLOPs needed to generate these results is

$(10^2 N\log N$ Flops/EF) x $(10^4 N \times M)$EF

where N is the number of atoms, EF is the number of energy and force evaluations, and M is the size of the coil subspace, typically 2-10 degrees of freedom. With future prospects for 100 teraflop resources, global optimization approaches can provide on the order of $10^6$ to $10^8$ low-energy folded conformations as well as high energy misfolded structures. Both provide an important feed-back loop on the quality of the energy surfaces, and therefore ultimately the means to quantitatively predict structures and fold proteins.

***Given the relevance of global optimization research to other scientific disciplines such as materials, combustion, and chemistry, it will be important to test both the scientific progress, and the parallel scalability of the implementation, on the protein structure refinement problem to fully exploit future high-end computing resources realized through the Strategic Simulation Initiative.***



***Global optimization prediction from sequence (right) and crystal structure (left) of the a-chain of Uteroglobin***. Methods like these can be adapted to use soft constraint directives predicted by fold recognition algorithms (Chapter 2) to refine protein structures.

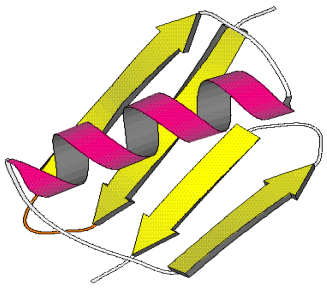
## Protein Folding Free Energy Landscapes from Free Energy Simulations

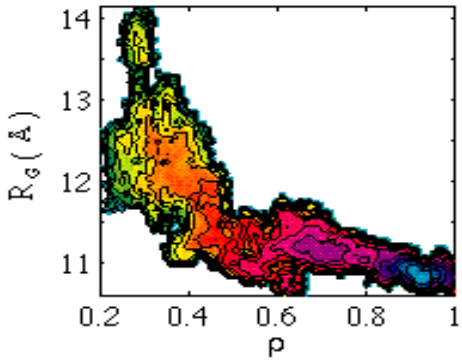
A critical test of atomic level models is mapping the free energy of a protein as a function of a small set of parameters that describe the protein folding reaction. Recent calculations have explored such folding free energy landscapes for proteins representing different topologies of the size that can be "predicted" from *de novo* folding experiments and readily identified by fold recognition algorithms.

Folding landscapes simulation studies have provided the general picture that α-helical proteins dominated by sequentially local interactions, fold via a mechanism that involves commensurate formation of tertiary and secondary structure, whereas β proteins must first collapse and then search through manifolds of collapsed states to "find" their native sets of secondary and tertiary structure interactions. These conclusions are manifest in the "diagonal" versus "L-shaped" folding free energy landscapes when projected onto coordinates describing overall collapse, such as the radius of gyration, and formation of native tertiary structure (Figure 1). Since the collapse-search mechanism is highly susceptible to kinetic traps, these calculations can aid aspects of protein design for β proteins by suggesting incorporation of strong turn formers.

The calculation of folding free energy landscapes for these proteins each required new advances in simulation techniques and developments of algorithms for massively parallel computers using partially distributed data message passing techniques. The computational cost of computing one such landscape is approximately one month of computer time on 512 processors of the Cray T3E-900, an architecture with a peak performance capability of about 0.5 teraflops. In addition to the raw CPU needs, 100 gigabits of data are produced and processed for each calculation. These calculations were still only possible with some truncation of the atomic model.

***The combination of CPU requirement and data storage and manipulation have meant that these calculations can only be carried out at computational facilities that possess the infrastructure to permit very long and relatively data intensive calculations to be performed. Advances in hardware and software support realized through the will allow full exploration of folding landscapes for all topological folds using better atomic models.***






***Protein topology and free energy landscape deduced from atomic level calculations.*** An α/β protein fold (left) shows an "L-shaped" free energy landscape (right) linked to a folding mechanism involving first collapse and then search for the native folded state among the multitude of collapsed states.

# Simulating the Function of Protein Kinases

Protein kinases are essential elements of most known pathways of signal transduction. One member of this large family is cAMP-dependent protein kinase (cAPK). cAPK has been demonstrated to be directly involved in cell cycle control and in the phosphorylation and activation of transcription factors, and is an important targets for new treatments for cancer and other diseases.

In vivo, cAPK exists as an inactive tetrameric protein that, upon binding cAMP, undergoes dissociation to a single dimeric regulatory subunit and two active catalytic subunits. The catalytic subunit is mainly found in three different conformations, in a closed conformation in ternary complexes with ATP or ADP and a peptide inhibitor, in an intermediate conformation in binary complexes with adenosine or balanol, and in an open conformation for the unliganded protein. Ligand binding therefore appears to induce hinge-bending closure of the two domains. Small molecules that can potently and selectively inhibit cAPK are of considerable interest as potential drugs.

Binding free energies of a natural inhibitor, balanol, and of a series of related inhibitors have been qualitatively reproduced by Poisson-Boltzmann continuum electrostatics calculations on static protein conformations. Similar calculations were used in an attempt to calculate the free energy for domain closure. Given the complexity of the internal interactions and dynamics of enzymes, it is likely that critical aspects of the molecular behavior is lost due to omission of entropic contributions of the protein (requiring many configurations), and to approximations in the non-atomistic models employed.

***To quantitatively determine free energies of drug binding and the long time scale motions of the hinge-bending closure of a protein like cAPK will require 100 TFLOPS computing performance, with proportional increases in system memory, to simulate microsecond molecular dynamics trajectories required to understand enzyme function.***

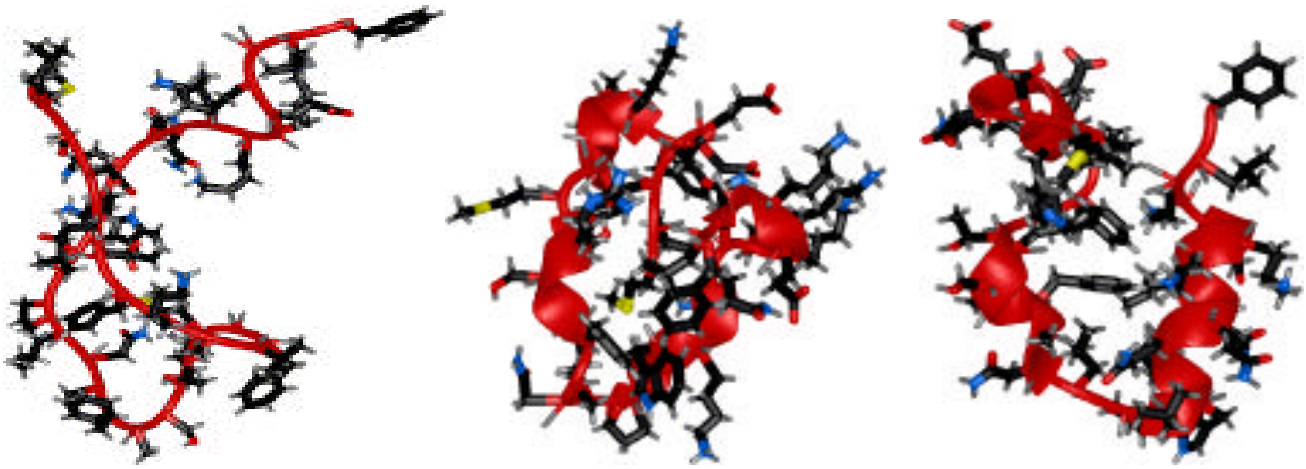

***Crystal structure of cAPK*** as reported by Susan Taylor and co-workers in Science, July 26, 1994**.** Knowledge of the cAPK structure allowed for the exploration of function for this protein using simulation. A major goal of the Accelerated Computational Biology effort is to determine structure and function directly from computation.

# Forward Folding Simulation of a Fragment of the Villin Protein

A 1 microsecond folding simulation of a fragment of villin protein that is stable to 70 degrees and is a compact, 3 helix bundle was carried out starting from an unfolded structure. Although the estimated folding time for this protein is 10-100 microseconds, the 1 microsecond simulation does reach a metastable structure that has many features of the native structure.

This trajectory, which is about two orders of magnitude longer than any previous of a protein in water, required the equivalent of about 2 and 1/2 months of dedicated time on a T3E computer. That work was catalyzed by some seed time on the Cray T3D at the Pittsburgh Supercomputer Center and continued on the T3E at SGI/Cray Research. Effective parallelization on the T3E architectures was achieved by spatial decomposition and separate treatment of water and protein parts of the system, but at the expense of not including long range electrostatic effects.

***The need for several orders of magnitude of computing speed beyond that available today is to do further interrogate empirical protein and aqueous force fields, provide a much better job of incorporating important long-range electrostatic effects, to simulate the smallest and fastest folding proteins, and give insight into the nature of protein folding.***



***One microsecond simulation of a fragment of the protein, Villin.*** The unfolded state from which the simulation started (left), a marginally stable state from simulation (center), the native NMR structure (right).

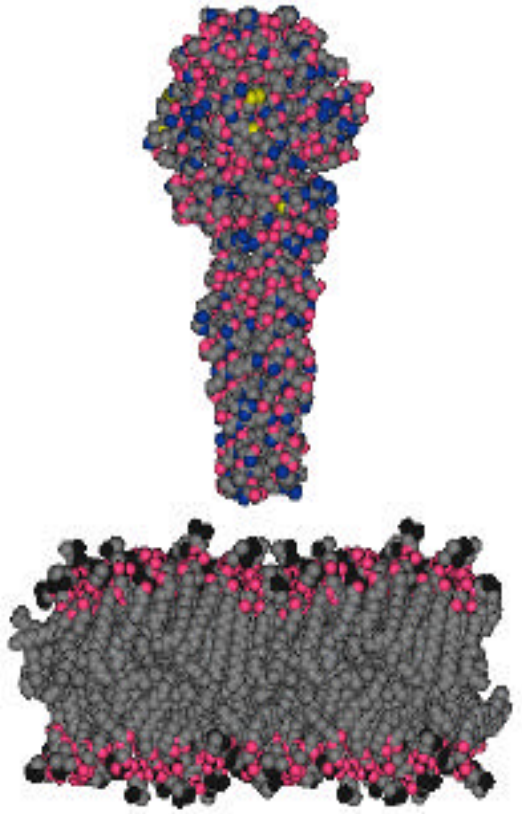
# Simulation of the Molecular Mechanism of Membrane Binding of Viral Envelope Proteins

The first step of infection by enveloped virus proteins involves the attachment of a viral envelope glycoprotein to the membrane of the host cell. This attachment leads to fusion of viral and host cell membranes, and subsequent deposition of viral genetic material into the cell. It is known that the envelope protein, hemagglutinin, exists under normal conditions in a "retracted" state where the peptide that actually binds to the sialylated cell surface receptor is buried some 100Å from the distal tip of the protein, and is therefore not capable of binding to the cell membrane. However, at low pH a major conformational change occurs whereby the fusion peptide is delivered (~100Å!) via a "spring-loaded mechanism" to the distal tip where it is available for binding to the cell membrane.

While the crystal structures of the inactive (neutral pH) and active (low pH) forms are known, the molecular details of the large-scale conformational change that produces the activated hemagglutinin, and of the binding of the fusion peptide to the cell surface receptors, are not known. As many aspects of this process are likely prototypical of a class of enveloped viruses including HIV, a detailed understanding of the molecular mechanism would be beneficial in guiding efforts to intervene before the viral infection is completed.

Atomistic simulations of the process of viral binding to cell membranes are extremely demanding computationally. Taken together, the hemagglutinin, lipid membrane, and sufficient water molecules to solvate the system in a periodically replicable box adds up to more than 100,000 atoms, an order of magnitude larger than what is presently considered a large biomolecular system. Furthermore, the large scale conformational changes in the protein, and displacements of lipids as the protein binds the membrane involve severe timescale bottlenecks as well.

***The combination of high end computing resources realized during the SSI, readily parallelizable smooth Particle Mesh Ewald with its NlogN scaling properties, and recent advances in MD algorithms such as multiple time step integrators and variable transformations, will be crucial in realizing mechanistic simulations of viral infection in the future.***



***X-ray crystal structure of the fusion-activated (low pH) form of the influenza virus hemagglutinin poised above a small (256 lipids) patch of a lipid membrane from an MD simulation***. The N-terminal peptides that actually bind to the membrane are not present in this hemagglutinin structure. An all-atom model of this complex plus solvent in a periodically replicated box would contain >100,000 atoms, and long timescale motions would need to be simulated to understand this step in the mechanism of viral infection.

## Models and Algorithms

In order for computational biologists to effectively employ teraflop computers, the empirical energy force fields, numerical algorithms, and software paradigms commonly used in the field must be improved. In this section, the developments in energy surfaces, algorithms and software support that will allow biological teraflop computing to realize prediction of biochemically relevant structures become a reality are discussed.

### *Atomic Energy Function Models*

The translation of protein sequence to protein structure rests upon a central dogma of biology: *proteins adopt their lowest free energy conformation as their functional state*. Thus, key to the success in any computational method that aims to provide sequence to structure predictions, or refinements, is that the energy function model used to represent the biological system yields the functional structure of known proteins as its lowest free energy state. Empirical protein force fields, which have formed the major component of all computational studies of protein structure, function and dynamics to date, give encouraging results in this regard. However, this conclusion is only qualitatively true for a handful of known proteins, and we need to explore the various ways in which protein surfaces can be modeled according to increasing levels of sophistication depending on the quantitative need.

Empirical protein force fields represent bonds and angles as harmonic distortions, dihedrals by a truncated Fourier series, and pairwise nonbonded interactions via Lennard-Jones 6-12 terms and Coulomb's Law for electrostatic interactions between point charges.

$$V_{MM} = \sum_{i}^{\#Bonds} k_b (b_i - b_o)^2 + \sum_{i}^{\#Angles} k_\theta (\theta_i - \theta_o)^2 + \sum_{i}^{\#Improper s} k_\tau (\tau_i - \tau_o)^2 + \sum_{i}^{\#dihedrals} k_\chi [1 + cos(n\chi + \delta)] + \sum_{i}^{\#atoms} \sum_{i<j}^{\#atoms} \frac{q_i q_j}{r_{ij}} + \varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \quad (3.1)$$

There are several protein force fields of this type in use, and it is clear that they are improving in their ability to represent protein conformations near the native fold, but have not been fully-tested outside of this local region on the energy surface. If water is included, and long range electrostatic effects with methods such as Particle Mesh Ewald are used, a simulation will conserve the protein backbone to within 1-2Å RMS of the crystal or NMR structure. This is roughly in experimental error since solution NMR and X-ray crystal structures often differ in backbone RMS by about 1Å. Some testing indicates that they do not always perform well outside of this local region, and therefore their usefulness in protein structure prediction and folding, which requires a good non-local description of the surface, is uncertain.

Beyond the empirical force fields for proteins is the problem of describing a solvent environment and its influence on the protein's conformational

behavior. The importance of hydration as a major contributor to protein stability and driving force for folding is widely accepted; in particular it is the hydrophobic interaction that is thought to be dominant, has originally pointed out by Kauzmann (1959). Simple models of hydration have been added to empirical protein force fields to attempt a better balance between computational cost and accuracy. One functional form of a simple model is to use a generalized Poisson Boltzmann treatment for the electrostatics, and to include a solvent-accessible surface area term to describe the free energy attributable to the hydrophobic effect. Another is to represent the solvent as a dielectric continuum rather than using explicit molecular water to capture the leading order contributions from electrostatics.

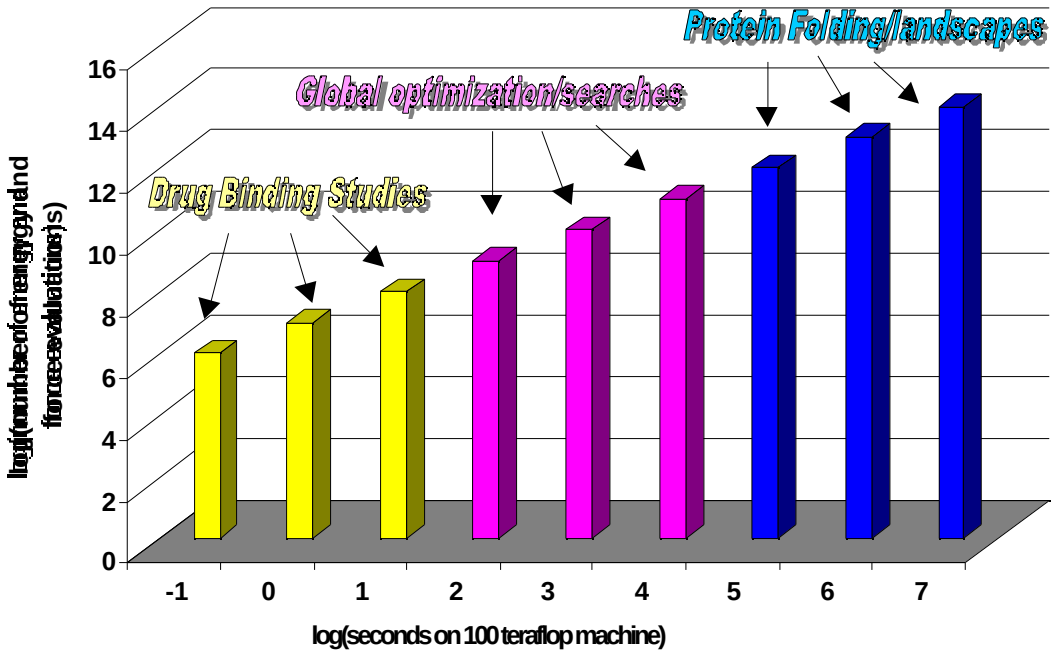
Development of a new set of implicit water potentials for biomolecular simulations is also an important direction. The purpose is to incorporate the statistical properties of water into solute-solute interactions and thereby avoid the computational limitations of simulating the polar solvent, which can be a major obstacle to biomolecular simulations in some cases. Success in developing the implicit solute-solute potentials, should lead to future peptide and protein simulations without explicit simulation of the water molecules, with their devastating spatial and temporal scales. Furthermore, eliminating the fast temporal scales of water dynamics is compatible with assuming constrained Langevin dynamics for the covalent bonds in the solute molecules. These implicit potentials can be fit to a convenient functional form and used in simulations of proteins to describe the important structural influence of aqueous hydration on protein conformations. They are physically complex but can be evaluated with reasonable computational cost, and are commensurate with both folding studies and protein structure prediction approaches using optimization, since they constitute a well-defined continuous force field, unlike the simpler descriptions above.

Explicit inclusion of molecular water is the most unambiguous way to describe a solvent environment around a protein, and has in fact been used in many molecular dynamics simulations. Empirical water force fields have been developed for the neat liquid over the last several decades, and more recently including explicit polarization, and do a quite reasonable job of reproducing a large number of molecular properties such as the partial radial distribution functions, thermodynamic and transport properties. While further improvements in the interface between water and protein force fields is warranted, explicit solvent calculations are important for quantitative studies of interaction of water with the protein.

***Smooth Particle Mesh Ewald is the method used to calculate the Coulombic energy and forces, the rate limiting kernal in classical MD simulations. For system sizes beyond $N=10^3$ atoms, these algorithms have largely reached their crossover to NlogN scaling. In addition, these Ewald-type calculations can be readily parallelized. The real space part of the sum can be treated using a standard domain decomposition strategy while the reciprocal space part of the sum requires the efficient parallelization of a three dimensional FFT with cutoffs in both reciprocal and real space.***

A fundamental difficulty encountered in the simulation of biomolecular systems is the need to evaluate long-range Coulombic forces, the first term in the double sum in Eq. (3.1). The evaluation of the Coulombic energy and Cartesian derivatives at a given protein configuration requires on the order of $N^2$ FLOPS, where N is the number of atom centers. The proper accounting of long-range forces is introduced through the Ewald summation. Typical protein in water simulations periodically replicate the system in three spatial dimensions, and divides the long range Coulombic interactions into a short range part that is evaluated in real space (as a direct sum over atomic positions) and a long range part evaluated in reciprocal space. Smooth Particle Mesh Ewald (SPME) employs a Cardinal B-spline based approximation to the atomic charge density, which can be generated using Fast Fourier Transforms, to calculate the Coulombic forces in reciprocal space in order NlogN.

For system sizes beyond $10^3$ atoms, these algorithms have largely reached their crossover to NlogN scaling. In addition, these Ewald-type calculations can be parallelized readily. The real space part of the sum can be treated using a standard domain decomposition strategy while the reciprocal space part of the sum requires the efficient parallelization of a three dimensional FFT with cutoffs in both reciprocal and real space.




***An estimate of the number of updates of Eq. (3.1) and its Cartesian derivatives that can be accomplished with NlogN scaling algorithms on a 100 teraflop computer for a 10,000 atom system (equivalent to a 100 amino acid protein and 3000 water molecules).***

The force field represented in Eq. (3.1) effectively incorporates many-bodied effects such as polarization effects through the parameters, but only requires the evaluation of two body forces. In particular, the molecular charge distribution is represented by partial charges of fixed magnitude assigned to atoms or sites, and the molecular charge distribution does not respond to the environment explicitly. More explicit inclusion of many-body polarization and many-body dispersion requires the evaluation of many-bodied forces, and can be accomplished through the use of a Drude model.

In a Drude model, the electronic degrees of freedom associated with a site/atom are treated by two non-interacting charges of opposite sign tethered together by a harmonic spring. The negative charge has a small mass, and the positive charge is fixed on the site/atom. Charges associated with different sites interact via Coulomb's Law. A classical treatment, minimization of the energy with respect to the light particle positions, yields many-body polarization up to dipole order. A quantum mechanical treatment yields dipole and higher many-body polarization as well as many-body dispersion. Modern path integral techniques can be used to simulate quantum Drude models efficiently. Finally, the Drude model scales like NlogN although the computational overhead is larger than a fixed charged treatment.

***Optimization and search strategies to construct biochemically relevant protein structures***

The quantitative determination of protein structure will be critical in extending the information that emerges from fold prediction to structures that are relevant for investigation of biochemical questions. It is well understood that current *de novo* and fold analysis techniques provide structural information that is “low resolution”, i.e., structures are typically 3-5 Å root-mean-square deviation from structural models emerging from NMR or X-ray structure determinations (these structures are typically precise to a level of 0.25Å - 0.75 Å). To extend the resolution of such structures to levels analogous to that from experimental structure determination methods, and hence to biochemically relevant levels, requires further refinement employing the types of force fields used in the folding free energy mapping calculations noted above. Only once we have achieved such resolution can we be confident in the use of these structural models as starting points drug discovery and functional assessment methods.

The problem of determining the full three-dimensional arrangement of the protein molecule in its most pragmatic guise is to ignore timescale bottlenecks for simulating the kinetics and mechanisms for how proteins fold, and instead determine effective ways of moving on the surface by walking through barriers. The conformational space of a protein is very high in dimensionality and complexity, so both local and global minima are of interest. The complexity of real protein surfaces rule out exhaustive enumeration of minima, so that sophisticated conformational searches and/or global optimization approaches are necessary to rapidly access the relevant regions of the energy surface. A large number of conformational search or optimization strategies have been developed to tackle protein structure prediction.

Simulated annealing, genetic algorithms, "non-local" dihedral angle Monte Carlo, and various mathematical optimization methods attempt to search more globally than local minimization algorithms. Simulated annealing is based on statistical mechanical theories for freezing in which the system is artificially "heated" to a high temperature and slowly cooled to "crystallize" to the lowest energy minimum. The correct cooling protocol and schedule is vital since a too rapid descent in temperature can result in trapping into metastable minima; advances in computing can allow the cooling rate to be a few orders of magnitude slower. Genetic algorithms define a set of genes composed of structural variables and their connection to a potential energy function; new genes are evolved by genetic crossover and random mutation, and genes that are unfit are eliminated from the population. Eventually, a population of genes (variables) is left, which in principle generates the lowest energy value. Non-local Monte Carlo methods have been developed where large moves are made with reasonably high acceptance when a small number of dihedral angles (backbone torsions $\varphi$ or $\psi$, or sidechain torsions) are varied according to probability maps of their amplitude derived from a representative set of proteins.

Mathematical optimization research is a more general approach for obtaining solutions to large nonlinear systems with numerous local minima, with protein folding being a recent example. Constrained optimization methods rely on the availability of sufficiently well-defined constraints so that the desired solution is the only available minimum, or one of few available minima, in the optimization phase of the algorithm.

***All global optimization approaches are exhaustive in the type of computational resources required due to the NP-complete number of minima arising from the complexity of the protein energy surface. However, it should be a tractable problem to provide on the order of $10^6$ to $10^8$ biochemically relevant structures from these methods, and many misfolded structures whose energies should be higher than the biochemically functional structures, using the techniques outlined in this section. This provides an important feed-back loop on the quality of the energy surfaces, and therefore ultimately the means to quantitatively predict and fold proteins from first principles either by simulation or optimization.***

Alternatively, global optimization and conformational search techniques attempt to systematically search the potential energy surface to find all low-lying minima including the global energy minimum.

The technique of "diffusion smoothing" is based on analogies to diffusion and heat conduction. A smoothing operator is applied to the potential energy surface to remove shallow minima and "absorb" them in non-shallow minima that become even deeper. Stochastic/perturbation is a global optimization algorithm that consists of two phases. In the first phase, a set of initial configurations is either designed or randomly generated, and each is used as a starting point for a local minimization. The best of the resulting local minimizers forms a pool used in the next phase. The second phase consists of repeatedly selecting conformations and modifying it using a small-dimensional global optimization probabilistic algorithm of

Rinnooy Kan. Various branch and bound methods have been reported for noble gas clusters and simple protein lattice model simulations.

***Simulation methodologies for dynamics and thermodynamics***

A dynamical description of how protein atoms and water molecules evolve in time can be determined by solving Newton's equations of motion

$$f_i = m_i a_i = - \ _i V(r_1, r_2, \ldots r_N)$$

A typical simulation is initiated by inserting a large biomolecule into a box of solvent, removing the overlapping solvent molecules, equilibrating and performing a long simulation. New positions and velocities are determined numerically using various finite difference algorithms,

$$r_i(t+\Delta t) = 2r_i(t) - r_i(t-\Delta t) + \frac{f_i(t)}{m_i}(\Delta t)^2 + O\left[(\Delta t)^4\right]$$

$$\boldsymbol{v}_i(t) = \frac{\boldsymbol{r}_i(t+\Delta t) - \boldsymbol{r}_i(t-\Delta t)}{2\Delta t} + O\left[(\Delta t)^3\right]$$

and the propagated error in the updated quantities are proportional to the power of the time step. Extended system equations of motion and associated numerical integrators have been developed that allow extensions from microcanonical ensemble dynamics to sampling of states in the canonical ensemble as well as the isothermal-isobaric ensembles.

The stability of these finite difference numerical integrator algorithms is dependent upon a time step that is commensurate with the fastest timescale in the system. Bond vibrations have an amplitude of 0.01Å and therefore limit the use of the central difference equations to time steps on the order of 1 femtosecond ($10^{-15}$ seconds). Constraint dynamics that effectively project out the force along bonds (SHAKE or RATTLE) can increase the time step to 2 femtoseconds (fs), so that the next fastest time scales arise from bond angle distortions. However, freezing bond angles has a significant adverse effect on developed structural and time scale properties of protein dynamics so that "Shaking" bond lengths and the 2 fs time scale resolution was part of the scaling behavior for simulations of protein-water and protein-protein interactions.

Recent advances in modern numerical integrators can now separate out the natural timescales of motions that depend on the strength of forces associated with each term in Eq. (3.1). Based on a factorization of the evolution operator used in quantum statistical mechanics, a formal decomposition of the integration time step allows bonds to be updated more frequently than angle bends, and angle bends more often than short range-forces, and short-range forces more often than long-ranged

***Recent advances in modern numerical integrators can now separate out the natural timescales of different motions in classical simulations. This formally correct multiple time step integration has been shown to generate an order of magnitude improvement in efficiency in simulating biomolecular systems. Calculations using multiple time step integration methods are very scalable; each time step results in a collective "move" and parallelization can proceed using standard domain decomposition paradigms.***

forces. This formally correct multiple time step integration has been shown to generate about an order of magnitude improvement in computational efficiency in biomolecular systems. The decrease in computer time results from the fact that the most expensive terms in Eq. (3.1), the double sum over atoms, need to be

updated less often than local interactions, i.e. the single sum terms in Eq. (3.1). Calculations performed using multiple time step integration methods in isothermal or isobaric-isothermal ensembles are very scalable. Each time step results in a collective "move" and parallelization can proceed using standard domain decomposition paradigms.

It is a difficult task to efficiently sample the conformational space of large complex single domain proteins. Large proteins relax on the second time scale in solution, a benchmark atomistic simulations cannot approach at present. It would be useful, therefore, to extended the practical range of polypeptide sizes which can be simulated and be said with reasonable confidence to have achieved conformational equilibrium. Although improved numerical integration and equations of motion have helped, several orders of magnitude improvement in efficiency needs to be obtained.

In Umbrella Sampling, a series of simulations are performed using not the true direct potential energy function but the true potential energy function plus a biasing potential. The biasing potential is characterized by a set of parameters that serve to adjust the strength of the bias along a "reaction coordinate". The biasing potential forces the dynamics to sample regions of the "reaction coordinate" that may not otherwise be explored extensively. Thus, a series of calculations at selected parameter sets can be performed to "drag" the system through its configuration space, along the reaction coordinate "in the shade of the umbrella" formed by the biasing potential. It is possible to achieve large reductions of computational effort if the "reaction coordinate", contains the rate limiting pathway(s) through configuration space. In general, Umbrella Sampling creates a simple level of parallelism because calculations employing a different set of biasing parameters can be performed independently. Postprocessing using the Weighted Histrogram Method can be used to eliminate the systematic bias in a formally exact manner.

The next level of complexity involves borrowing methods from the path integrals molecular dynamics literature. Specifically non-canonical, order N variable transformations increase the sampling efficiency of Gaussian random coil calculations by a factor of over two hundred. Most of the increase in efficiency can be ascribed to the non-canonical variable transformation that permits the long wavelength fluctuations of the coil to occur on a fast time scale. Applying such transformations with umbrella sampling has allowed the efficient and accurate determination of the hinge bending free energy surface of the mutant T4 lysozyme {in vacuo} and in computer water solution.

Finally, proteins have a much more complex configuration space than random coils, hinge bending modes or intermediate size peptides. It is therefore useful to implement yet another class of variable transformation borrowed from the Monte Carlo literature that can, in principle, make true protein/polypeptides "resemble" a random coil model by analytically eliminating torsional barriers along the peptide backbone. The fast random coil methodology can then be applied "on top" of this first transformation. The idea behind new the technique is to create through the use of noncanonical variable transformations a smooth effective energy landscape without a concomitant modification of the potential energy surface, itself. In contrast, standard torsional dynamics schemes seek to eliminate motion in

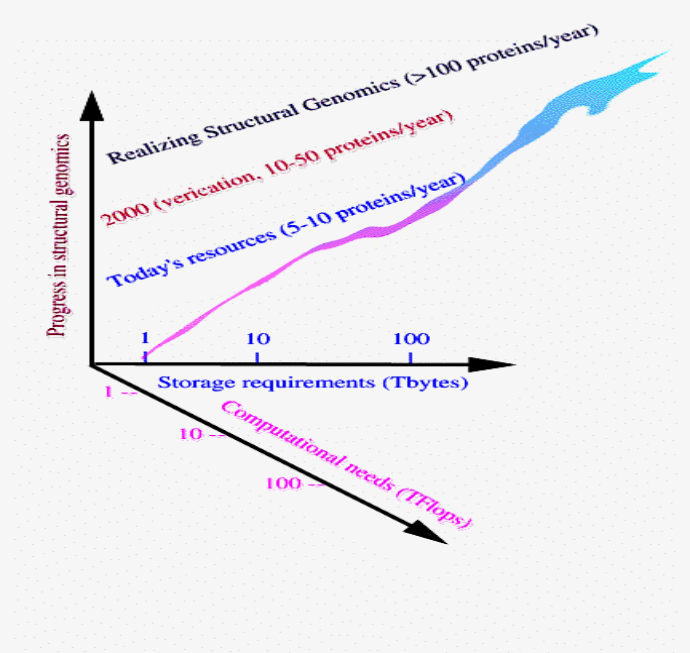
directions tangential to the backbone dihedrals to promote the use of larger time steps and barrier crossing events but do alter the heights of the barriers in the coordinate space. The new method should also be contrasted to importance sampling schemes where barriers are cut by an ad hoc modification of the potential energy surface and must be "regrown" by a {a posteriori} reweighting of the trajectories. In the new method, the reweighting occurs, dynamically, through the properties of the non-canonical variable transformation. Preliminary results are promising.

# High End Computing for Protein Structural and Functional Studies

We have illustrated in this Chapter that we are poised to address the milestones outlined for biomolecular simulations in the introduction. The effort possible from a Strategic Simulation Initiative of 100 tera-scale means that time and size scales become accessible for the first time for predicting protein structure and folding, and simulating protein function and thermodynamics.

- ***The simulation of a protein folding trajectory of a 100 amino acid protein and 3000 water molecules would require $10^{12}$-$10^{15}$ updates of Eq. (3.1) and its derivatives. Many folding trajectories will likely be needed to understand the process on how proteins self-assemble.***
- ***Unfolding to pathway intermediates such as the "molten globule" is thought to take on the order of a millisecond or $10^{12}$ steps; assuming microscopic reversibility, events early in unfolding are the same as late folding events that will give insight into the forward folding problem.***
- ***Optimization approaches for generating a diversity in biochemically relevant structures and misfolded conformers will take on the order of $10^{8}$-$10^{10}$ updates depending on the number of conformations and improvements in algorithms. Provides an important feedback loop for improving energy functions.***
- ***The simulations of relative free energies important for determining drug binding affinities can take on the order of $10^{6}$ steps, and the thermodynamic derivatives (entropy and enthalpy) >$10^{6}$ updates.***

To meet this challenge we must begin preparations now. Clearly, the resources required to yield data for just a few protein sequences represents the outer envelope of what is possible today. The Figure below outlines a tentative scenario of the computational and infrastructure "cost" associated with meeting the future challenge of predicting biochemically relevant structures from a structural genomics effort of scale.




***Computational cost of success in computational biotechnology.*** As suggested in this scenario, we have outstripped the resources available to the computational biology community with even the initial exploratory calculations discussed in this chapter.

Computational resources will need to be increased by orders of magnitude if the approaches being used today are to move beyond the realm of benchmarks and become the essential components of protein sequence to biochemically relevant protein structure to understand biological function and disease.